МИЛЛИЙ КОРПУСГА ҚЎЙИЛАДИГАН ЗАРУРИЙ ЛИНГВИСТИК ВА ЛИНГВОДИДАКТИК ТАЛАБЛАР

Мадина Шухратовна НОРБЕКОВА
таянч докторант
Алишер Навоий номидаги
ўзбек тили ва адабиёти университети
Тошкент, Ўзбекистон
madinabonushuhratovna@gmail.com

Аннотация

Мақолада миллий корпусга қўйиладиган лингвистик ва лингводидактик талаблар, миллий корпуснинг зарурати ҳамда унинг тил ўрганишдаги алоҳида аҳамияти ҳақидаги фикрлар ёритилган.

Таянч сўзлар: корпус, миллий корпус, электрон луғат, лингвистик ва лингводидактик, технологик талаблар, конкорданс**.**

НЕОБХОДИМЫЕ ЛИНГВИСТИЧЕСКИЕ И ЛИНГВОДИДАКТИЧЕСКИЕ ТРЕБОВАНИЯ К НАЦИОНАЛЬНОМУ КОРПУСУ

Мадина Шухратовна НОРБЕКОВА
базовый докторант
Университет узбекского языка и литературы
имени Алишера Навои
Ташкент, Узбекистан
madinabonushuhratovna@gmail.com

Аннотация

В статье выделяются лингвистические и лингводидактические требования к национальному корпусу, необходимость национального корпуса и его особая значимость в изучении языков.

Ключевые слова: корпус, национальный корпус, электронный словарь, лингводидактический, технологические требования, конкордантность.

Жаҳон тилшунослигида сунъий интеллектнинг автоматик таржима, компьютер таҳлили, таҳрири, тезаурус, электрон луғат сингари имкониятлари кенгайди, илмий-назарий асослари яратилди, амалиётда ишлатиш мумкин бўлган илк намуналари қўлланила бошлади. Дунё тилшунослигида корпус соҳасидаги мақсадли тадқиқотлар XX асрнинг қирқинчи йилларида Блумфилд, Фриз ва Бонджерслар томонидан бошланган. Н.Френсис ва Г.Кучера эса илк марта корпус тузиш принципларини ишлаб чиққан. Рус тилшунослигида В.П.Захаров, А.Б.Кутузов, Э.В.Недошивина,

В.В.Риков, В.А.Плунгянлар корпус, унинг турлари, корпус тузиш ва теглаш тамойиллари борасида тадқиқот олиб боришган. Корпус бирликларини лексик-семантик теглаш муаммолари Г.И.Кустова, О.Н.Ляшевская, Э.В.Падучева, Э.В.Рахилина, Ю.Д.Апресян, Л.Л.Иомдин, А.В.Санников, В.Г.Сизов тадқиқотлари предмети бўлган. Ўзбек тилшунослигида компьютер лингвистикаси, матнга лексикографик ишлов бериш ва лингвостатистик таҳлил этиш борасида муайян тадқиқотлар амалга оширилган. А.Пўлатов, С.Муҳаммедов, Н.Айимбетов, С.Муҳамедова, С.Каримов, Г.Жуманазарова, А.Бабанаров, Д.Ўринбоева, Ш.Ҳамроева, А.Норов ва бошқаларнинг изланишларини ана шундай ишлар сифатида қайд этиш ўринли. Фандаги бу янгиланишлар ахборот технологияларини тилшунослик ва таълимга татбиқ этишни кун тартибига қўйди.

Ўтган асрнинг олтмишинчи йилларида мазкур жараён жадаллашди, XXI аср бошларида ўзида миллионлаб сўзларни акс эттирувчи юзлаб тил корпуслари пайдо бўлди. Фандаги бу янгиланишлар ахборот технологияларини тилшуносликка татбиқ этиш билан боғлиқ истиқболли илмий йўналишлар пайдо бўлишига йўл очди. Бу эса корпус, корпус лингвистикаси, унинг шаклланиши, тараққиёти, бугунги ҳолати ва корпус тузишнинг умумий тамойилларини ўрганиш заруратини белгилайди. Такомиллашиб бораётган компьютер лингвистикаси йўналишида автоматик таржима сифатини яхшилаш, тилни лингвистик моделлаштириш, ҳар бир тилга оид сўзларни леммалаш назарияси, алгоритмини яратиш ҳамда муайян тилнинг кўп асрлик миллий-маданий меросдан фойдаланиш имконини ошириш мақсадида уларни электронлаштириш жаҳон тилшунослигида долзарб масалага айланди. Тилшуносликда, хусусан, компьютер лингвистикаси соҳасида корпус яратиш, мавжуд корпуслар ҳажмини кенгайтириш, матнни автоматик қайта ишлайдиган дастурларни ишлаб чиқиш кабилар ўз ечимини кутаётган муҳим масалалардан бири бўлиб турибди. Ўзбек тили миллий корпусининг лингвистик таъминоти сифатида Т.Нафасов, В.Нафасовалар томонидан тузилган “Ўзбек тили

топонимларининг ўқув изоҳли луғати”ни ҳам қайд этиш мумкин. Тадқиқотчи Ш.Ҳамроева “Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари” номли диссертациясида семантик теглаш ҳақида умумий маълумот берган бўлса, Д.Ахмедова тадқиқотида ўзбек тили атов бирликларини семантик теглаш масаласи махсус тадқиқ қилинган, ўзбек тили атов бирликларининг маълумотлар базасини ишлаб чиқишнинг айрим муаммолари ҳам ўрганилган. Масалан, А.Ешмўминов ўзбек тили миллий корпусининг синонимлар базасини ишлаб чиқиш масаласини ечишга ҳаракат қилган бўлса, Г.Бегматова ўзбек тили идиомаларининг миллий корпусда берилиши ҳамда унинг маълумотлар базасини ишлаб чиқиш муаммосини тадқиқ этган. Миллий корпус учун географик терминларнинг “Географик терминлар” базаси яратилиб, уларга семантик теглар бириктирилади. Шу параметрлар асосида миллий корпус семантик қидируви учун қидирув белгилари аниқланади. Бу эса миллий корпусда терминологик бирликларнинг жойлаштирилиши, уларни қайта ишлаш, қидирув имкониятлари ва материалларнинг бой тарздаги таъминоти учун хизмат қиладиган асосий омиллардан саналади [5]. Бугунги кунда кўплаб тил корпуслари мавжуд бўлиб, турли мақсадлар доирасида яратилганлиги боис улар тил корпусларининг махсус турлари ва турли имкониятларини намоён этади. Тил корпуслари бўйича тадқиқотлар қамрови кенг бўлиб, улар, асосан, турли мақсаддаги тил корпусларининг лингвистик ва дастурий таъминотларини яратишга, мавжуд корпусларни ривожлантиришга ҳисса қўшади.

Тил корпуслари, айниқса, замонавий электрон луғатлар тузиш ва ундан фойдаланиш маданиятини шакллантириш тил имкониятини эгаллашда самарадор эканлиги ўз исботини топган. Айнан тил таълими ва электрон луғатларни яратишда дунё миқиёсида яратилишига эҳтиёж ниҳоятда юқори бўлган тил корпусларининг ўрни беқиёс [3]. Маълумки, компьютер технологияларининг корпус тилшунослигида қўлланилиши натижасида тил корпуслари асосида сўзнинг қўлланиш даври ва частотасини аниқлаш,

терминология соҳасини ривожлантириш, гап қурилишини ўрганиш ва таҳлил қилиш, таржима дастурлари учун тилли база яратиш, услубиятни ўрганиш, электрон лингводидактикани ривожлантириш, айниқса, турли луғатларни яратиш имкониятларининг картотекадан автомат жараёнига ўтилди ва бемисл қулайликларни юзага келтирди [6].

Кўп мақсадли ва кўп вазифаларни бажара оладиган тил корпуси – муайян бир тилнинг Миллий корпуси ҳисобланади.

Миллий корпус қуйидаги талабларга жавоб бериши зарур:

1) тилнинг барча услубларига мансуб ўта катта ҳажмдаги матнларнинг бўлиши;

2) матнларнинг лингвистик, морфологик ва синтактик жиҳатдан тегланиши;

3) корпуснинг мета-маълумотлари мавжуд бўлиши [3].

Мазкур талаблар миллий корпуснинг зарурий лингвистик ва лингводидактик технологик воситага айланишини таъминлайди.

Замонавий корпуслар – бу маълум бир тилда, электрон шаклдаги матнлар тўпламига асосланган ахборот-маълумот тизимидир. Ҳар бир корпус филологик ёндашув билан матнлар устида ишлаш учун, албатта, лингвистик аппарат ва дастурий таъминот билан таъминланиши жоиз. Бугунги кунда корпуслар луғатлар ва грамматика каби тилшуносликнинг ажралмас қисмига айланди. Корпус пайдо бўлганидан сўнг тилшунослик фанлари ўзгариб кетди, айтиш жоизки, бутун тилшунослик корпус тилшунослигига айланди. Энг таниқли ва тан олинган лингвистик корпусларга намуна сифатида қуйидагиларни келтириш мумкин: Рус миллий корпуси (хттпс://руссорпора.ру/нew/), Британия миллий корпуси (хттп://www.натсорп.ох.ас.ук/, хттпс://www.енглиш-сорпора.орг/бнс/), Турк миллий корпуси (хттпс://www.тнс.орг.тр/), Америка миллий корпуси (хттп://www.анс.орг/) ва бошқалар. Табиийки, бундай корпуслардан жуда кўпларини кўришимиз мумкин. Шуни ёдда тутиш керакки, баъзи бир йирик корпуслар маълум бир вазифаларни бажарадиган, фойдаланувчиларнинг

маълум бир доирасига йўналтирилган бир нечта махсус кичик корпуслар ёки қисмий корпусларни ўз ичига олиши мумкин. Масалан, рус миллий корпусининг асосий таркибида 10 та қўшимча махсус қисмий корпуслар бор. Мавжуд корпусларни таҳлил қилиш натижасида кўриш мумкинки, корпусларни қуришда барча тиллар учун ягона бўлган методология мавжуд эмас [3]. Бунинг сабаби шундаки, турли тилларда турли хил қоидалар, технологик жараёнлар амал қилади. Морфологик характеристикалар тўпламлари турлича бўлишига қарамай, юқорида қўриб чиқилган корпусларнинг барчасида морфологик (ёки грамматик) разметка таъминланган. Синтаксис разметка эса баъзи корпусларда бор, лекин турлича ёндашувлар асосида амалга оширилган бўлса, баъзиларида умуман йўқ. Корпусларни тилни қамраб олиш даражаси бўйича солиштирганда, рус тили миллий корпусини ҳам хронология, ҳам жанрлараро энг мувозанатлашган ва матнлар хилма-хиллиги таъминланган корпус деб ҳисоблаш мумкин. Лекин рус тили миллий корпусининг бу ютуғи унинг асосий камчилигини ҳам юзага келтирган. Ҳар қандай йирик корпусда тўлиқ ва аниқ разметкалашнинг имконияти чекланган бўлади. Ноаниқликлар, асосан, автоматик разметкалашда омоним сўзлар туфайли пайдо бўлади. Корпус лингвистикаси бўйича тадқиқотлар шуни кўрсатадики, тил тадқиқотларини ўрганиш учун мос асос бўлган машинада ўқиладиган матнлар тўплами билан шуғулланади. Матнлар тўплами, одатда, ҳар қандай оқилона вақт оралиғида фақат қўл ва кўз билан таҳлил қилишга тўсқинлик қиладиган ҳажмга эга. Шу сабабли, корпуслар доимо дастурий таъминотнинг қидириш воситаларидан фойдаланади. Конкордан фойдаланувчиларга сўзларни контекстда кўриш имконини беради. Бошқа воситалар частота маълумотларини ишлаб чиқаришга имкон беради. Масалан, сўзлар частотаси рўйхати, унда корпусда пайдо бўладиган барча сўзлар рўйхати ва уларнинг ҳар бири ушбу корпусда неча марта содир бўлишини белгилайди. Мувофиқликлар ва частота маълумотлари корпус лингвистикаси учун бир хил даражада муҳим бўлган таҳлилнинг иккита шаклини, яъни сифат ва миқдорини кўрсатади [7].

Мавжуд корпуслар таҳлилидан кўриш мумкинки, миллий корпус тилнинг қонуният ва хусусиятларини иложи борича кўп тақдим эта олиши ва таркиби бўйича мувозанатлашган бўлиши керак. Матнларнинг бир тури тилдаги жами матнлар ичида қандай ҳажмий улушга эга бўлса, ана шундай миқдор миллий корпусда ҳам сақланиб қолиши керак. Лингвистик корпуслар қўлланилишига қуйидаги талаблар қўйилиши мумкин [1]:

1) корпус ўқув ва маълумот қўлланмалар ёзишда, луғатлар тузишда ҳамда лингвистик экспертизалар ва таҳлилларни ишлаб чиқишда, тил маълумотлари сифатида ишончли ва ваколатли манба ҳисобланади;

2) корпус турли хил лингвистик фанларни ўқитишда худди тилнинг кўргазмали (иллюстрация) манбаи ҳамда таҳлилларни тақдимот сифатида ишлатиш мумкин;

3) корпус илмий тадқиқот ишларида, шу ўринда илмий фаразларни текширувчи сифатида қўлланилиши мумкин;

4) корпус тил фаолиятининг универсал маълумотлар қўлланмаси сифатида бўлиши мумкин.

Ўзбек тили корпусини ишлаб чиқиш бугунги кунда эътироф этилган стандартлар асосида амалга оширилади. Сўз бирикмаларининг боғлиқлигини ва уларнинг хатоликларини аниқловчи алгоритмлар таҳлил этилиши муҳим масала ҳисобланади. Корпуснинг қидирув тизими сўзлар, сўз бирикмалари, лемма, токен ҳамда н-грам моделлининг конкорданс қидирув тизими ва дастурини ишлаб чиқиш билан баҳоланади. Миллий тил корпусининг пайдо бўлиши унинг инсоният ҳали корпус лингвистикаси фани пайдо бўлмасдан олдинги даврда, яъни XX асрдан бошлаб тадқиқ этила бошланган. Библияни тадқиқ этиш ҳамда луғатлар яратиш [8], тилларни ўқитиш, дескриптив грамматика ва бошқалар таҳлил этилган. Квирк корпуси бир миллион сўз бирикмаларини ўзида жамлаган бўлиб, ҳар бири ўн етти қатор матндан иборат миллион картотекадан иборат. Мазкур корпуснинг яратилишига 25 йил вақт сарфланган ва Квирк корпусининг ишлари 1989 йил охирида якунланган. Бу пайтда технологиялар юқори суръатда ривожланганлиги боис

корпус тезликда электрон шаклга ўтказилган. Маълумотларда аниқлилик муаммоси туфайли тилни равонлаштириш н- грамм тил модели муҳим техник алгоритмдир. Бироқ, баъзи ҳолларда у кўринмас н- граммларга нотўғри эҳтимоллик миқдорини белгилайди. Шунинг учун ушбу мақолада тилнинг агглютинатив хусусиятларидан фойдаланган ҳолда кўринмас граммларнинг нотўғри тайинланган эҳтимолларини мослаштирадиган янги усули тақдим этилади. Тилда а морфеманинг грамматик тоифасини олдинги морфемани билиш орқали олдиндан айтиш мумкин. Ушбу белгидан фойдаланиб, грамматик жиҳатдан нотўғри бўлган н-граммаларнинг нисбатан юқори эҳтимолликка эришишига йўл қўймасликка ҳаракат қилинади.

Миллий корпус қандай ривожланади? Рус тилининг миллий корпуси, аввало, XIX асрнинг ўрталаридан XXI асрнинг бошларигача бўлган даврни ўз ичига қамраб олди. Бу давр хоҳ ўтган, хоҳ янги бўлишидан қатъи назар, у социолингвистик кўринишдаги бадиий сўзлашув, жонли сўзлашув, қисман матнларни ташкил қилади. Корпусга бадиий қимматга эга бўлган ва тил ўргатишга қизиқиш уйғотадиган бадиий адабиёт намуналари киритилади. Бадиий шеърлардан ташқари, ёзма адабиётнинг бошқа, ёзма адабиётнинг намуналаридан публицистика, илмий-оммабоп ва илмий адабиётлар, шахсий чиқишлар (маърузалар), шахсий ёзишмалар, кундаликлар, ҳужжатлар киритилади. Миллий корпуснинг ўзига хос муҳим хусусияти меъёрлаштирилган муайян таркибга эга эканлиги билан характерланади. Бу корпус маълум тилда берилган (турли бадиий жанрлар: публицистик, ўқув, илмий, иш юритиш, сўзлашув, публицистик, сўзлашув, шевавий каби), уларнинг барчаси имкон даражасида маълум доирага оид маълумотларнинг пропорционал матнлар ҳисобланадиган оғзаки ва ёзма кўринишларининг барчасини ўз ичига олади.

Демак, корпуснинг қониқарли даражада бўлиши учун унинг кўламига эътибор қаратиш кераклигини назардан четда қолдирмаслик керак (масалан, ўн ва юз миллионгача сўз қўллаш).

ФОЙДАЛАНИЛГАН АДАБИЁТЛАР:

1. Баранов А. Н. Корпусная лингвистика // Баранов А.Н. Введение в прикладную лингвистику. – Москва: Эдиториал УРСС, 2001. — 360 с.
2. Захаров В.П. Корпусная лингвистика. – Иркутск: СПбГУ. 2005. – 48 с.
3. Kompyuter lingvistikasi: muammolar, yechim, istiqbollar: xalq. Ilm.amal.konf. – Тошкент: – Alisher Navoiy nomidagi TDO‘TAU, 2022. – №. 01. – 289 b.
4. Raupova L. O‘z qatlamga doir grammatik terminlar va ularning elektron lug‘atini yaratish muammolari //Kompyuter lingvistikasi: muammolar, yechim, istiqbollar: xalq. Ilm.ama.konf. – Тошкент, 2022. – №1. – 289 b.
5. O‘zbek tilining milliy korpusi: muammo va vazifalar: xalq.ilm.amal. konf. – Toshkent: TDO‘TAU, 2022. – №. 01. – 352 b.
6. O‘zbek tili milliy korpusi – muhim madaniy voqelik (davra suhbati). //Ma’rifat, 2021. – 11.08. – № 32 (9357).
7. McEnery T, Wilson A. Corpus Linguistics. Edinburgh: Edinburgh University Press, 2nd edition, 2001.
8. S. Chen, D. Beeferman, and R. Rosenfeld. Evaluation metrics for language models. Proceedings of the DARPA Broadcast News Transcription and Understanding Work-shop, pages 275–280, 1998.