

КОРПУС ТУЗИШНИНГ АЙРИМ ҚОИДАЛАРИ ҲАҚИДА ABOUT SOME CONSTRUCTION RULES

Гули Тоирова*

Аннотация: Мақолада корпуснинг қўп қиррали лингвистик манба сифатидаги асосий аҳамияти, корпус асосан икки турдаги маълумотга эга эканлиги ҳамда унинг турлари таҳлил қилинган. Лексик-грамматик кодни шакллантириш мақсадида фонологик, морфонологик ва орфографик қоидалар алгоритми зарурияти илмий асосланган. “Эксперт лингвистик тизим” корпус-менежер тизими ва унинг асосий имкониятлари айтиб ўтилган. Ўзбек тили миллий корпусининг лингвистик базасини тузишда фонологик, морфонологик ва орфографик каби лингвистик модулларнинг аҳамияти, корпусда асосий структураларни бирликлари ажратилиши ёритилган.

Калим сўзлар: корпус, катта массивли матн маълумоти, лексик маълумот, морфологик белги, сўз, gap, абзац, матн, разметка.

Annotation: The article analyzes the main role of the corpus as a multifaceted linguistic source, the fact that the corpus basically has two types of information, and its types. The need for an algorithm for phonological, morphological and spelling rules for the formation of the lexical and grammatical code is scientifically substantiated. The corpus management system "Expert linguistic system" and its main features are described.

Keywords: corpus, large array of textual information, lexical information, morphological sign, word, sentence, paragraph, text, mark.

Ҳозирги жамиятда инсонлар ўртасидаги мулоқот жараёнлари янада мураккаблашиб, мукаммаллашиб бормоқда. Компьютер пайдо бўлиши билан мулоқотнинг интернет шакли юзага келиши кишилик жамияти учун қай даражада қулайликларни яратаётган бўлса, ўзгараётган шароитларда инсон дунёни табиий қабул қилиш қўнқмаслиги ва она тилини ўз миллий-маданий анъаналари доирасида сақлаб қолиши зарурлигини тақозо қилмоқда. Зеро, улкан матн маълумотларини тезкор қайта ишлаш имконига эга бўлган техниканинг пайдо бўлиши олам лисоний манзарасини вертуал оламга қўчирмоқда.

Компьютер технологиялари воситасида ахборотни қайта ишлаш, машина таржимаси, электрон лугатшуносликни ривожлантириш, тезауруслар тузиш, тил корпусини яратиш ишларини дунё лингвистикаси цивилизациясига эришган тиллар аллақачон амалга оширганлигини таъкидлаган эдик. Инглиз, рус, араб, француз, немис, испан, тожик каби тиллар шулар жумласидандир. Мазкур тилларда интернет тизимида тил корпусини яратиш масаласининг илмий-назарий жиҳатлари ҳам ишлаб чиқилганки, бу ўзбек тилини ҳам интернет “тушунадиган” тилга айлантириш (яъни тил корпусини яратиш) сабъ-ҳаракатларини жадаллаштириш заруратини кун тартибига қўяди.

Бу борада тилшунос олимлар тилни ва у билан боғлиқ бўлган ҳамда илгари маълум бўлмаган соҳаларни ўрганишда жуда муҳим тадқиқот имкониятига эга бўлдилар [Н.З.Абдурахмонова 2018:49; Абжалова М. 2019:49; Ҳамроева Ш. 2018:52 ; Эшмўминов А. 2019: 45]. Тилни ўрганишда тилшунослар учун нутқ фаолиятини анча тўлиқ акс эттирадиган қулай ва ихчам нутқ материалига эҳтиёж сезади. Бу эҳтиёж эса электрон шаклда тақдим этилган матнларни, яъни лингвистик корпусларни яратишни тақозо қилади. Лингвистик корпуслар ўзида тил тизимининг тилшунослик изоҳларини ва бошқа тилшунослик фанларида қўлланилиши мумкин бўлган матнларни корпусда ўзига хос акс этадиган нутқ материали сифатида тадқиқ этади. Анъанавий тилшуносликдан фарқли равишда корпус тилшунослиги тил ёки нутқ ходисаларини ўрганиш билан шуғулланмайди, балки нутқнинг корпусда махсус танланган матнлари барча қўринишларини (бадий, илмий, муомала тилини) ўрганади. Бу ўринда мантикий фикрлаш ва анъанавий тилшунослик дедукциясига нисбатан эҳтимоллик, статистик ва индуктив

^{36*}Филология фанлари бўйича фалсафа доктори (PhD), доцент. Бухоро давлат университети.
tugulijon@mail.ru

усуллар тез қўлланилади. Бунда узоқ йиллар давомида йиғилган тажрибалар асосида олинган натижаларга, баҳоларга ва амалларга эмас, балки эмпирик жиҳатдан тўпланган материалларга кўпроқ таянилади. Корпус тилшунослиги учун асосий восита бир ёки бир неча тиллар доирасидаги нутқ фаолияти ҳамда корпусдаги кейинги тадқиқотларни ихчамлаштириш учун зарур бўлган корпус материалларини кодлаштириш ҳисобланади. Қуйидаги хусусиятлар корпус учун муҳим омил саналади: тўпланган материалларни, албатта, машина сақловчиларига (компьютерлар хотирасига) жойлаштириш; электрон қидирувга (морфологик, синтактик қидирувлар) имкон берадиган ўзига хос белги; якуний ўлчов бирлигига эга ва репрезентативлик (тилдаги кўплаб жанрларнинг асл ҳолида тўлиқ акс этиш).

Ўзбек тили миллий корпуси – ўзбек тилининг бадий асарлари ҳамда вариантлари, кенг қўламли лингвистик ва метаматнли маълумотлари билан таъминланган, турли хил услубларни акс эттирувчи электрон матнлар мажмуасини ифодалай олиши керак.

Бундай маълумотлар мавжудлиги корпуснинг асосий хусусиятларидан бири бўлиб, уни электрон шаклдаги ҳаммабоп оддий матнлар мажмуидан, хусусан одатдаги интернетдан фарқли ўлароқ ажралиб туради. Маълумотнинг мукамаллиги ва аниқлиги, шунингдек, турли хилдаги тилга оид факт ва ҳодисларнинг батафсил қамраб олиниши корпуснинг кўп қиррали лингвистик манба сифатидаги асосий аҳамиятини касб этади.

Корпус асосан икки турдаги маълумотга эга бўлади:

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

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

В.1. Морфологик белгилар:

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

В.2. Семантик белгилари:

семантик разряд, таксономик синф, мереология, баҳо, каузация, сўз ясовчи алоқалар ва б. [Аброскин А. А. 2009:277-282 ; Касьянов В. Н., Касьянова Е.В. <http://pco.iis.nsk.su/ICP>; Касьянова Е.В. 2004:189–205; Ҳамроева Ш. 2018:52].

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

Шундай қилиб, корпусда асосий структурали бирликларни ажратиш мумкин: сўз, гап, абзац, матн. Бунда матнда структурали бўлиниш (қисмлар, боблар, бўлимлар)ни ифодаладиган, абзацдан ташқарида бўлган бирликлар ва гапнинг синтаксик структураси (клауз, гуруҳлар)ни ифодаладиган бирликлар ишлатилмайди. Маълумот фақат структуранинг минимал ва максимал бирликларига ёзилади: сўзга ва бутун бир матнга. Синтактик белгилар (разметка)нинг баъзи бир элементлари турғун сўз бирикмалар (иборалар)ни белгилаш учун қўлланилади[Касьянова Е.В. 2004:189–205.].

Белгининг ҳажми ва мукамаллиги жиҳатидан корпус иккита тенг бўлмаган қисмга бўлинади. Корпуснинг асосий қисмини матнлар ташкил этиб, унда ҳар бир сўзга автоматик равишда турли-туман морфологик таҳлиллар ёзиб қўйилади. Корпуснинг маълум бир қисмида мукамалроқ белги (разметка) ўз аксини топган, аниқроғи: морфологик омонимия қўлда бажарилиб олиб ташланган, сўзларга семантик белгилар ёзиб қўйилган, шунингдек, сўзларга урғу қўйиб чиқилган. Корпуснинг мазкур қисми аниқ маълумотни беради ва лингвистик корпуслар учун эталон вазифасида ўташи ҳам мумкин. Метаматнли белги (разметка) корпуснинг иккала қисмида бир хил характерга эга бўлади[Аброскин А. А. 2009:277-282].

Ҳозирги вақтда корпус матнларини қайта ишлаш учун зарур бўлган дастурлар етарли даражада мавжуд. Масалан, «Эксперт лингвистик тизим» корпус-менежер тизими шундай дастурий тизимлардандир. Унинг асосий имкониятлари қуйидагилардан иборатдир:

- 1) танланган матнлар асосида лексема ва сўзшаклларининг такрорланиши луғатини яратиш имконияти;
- 2) олинган луғатининг ҳар қандай бирлиги учун матнни кўриб чиқиш имконияти;
- 3) графикли сўзни бўғинга ажратиш;
- 4) сўз заҳираларини саралаш;
- 5) бир вақтнинг ўзида чекланмаган файлларни қайта ишлаш имконияти;
- 6) ташқи белгиларга эга бўлган матнлар корпусларини яратиш имконияти;
- 7) яратиладиган матнлар корпуслари ҳамда корпусга кирувчи алоҳида матнлар учун статистик маълумотларни ҳисоблаб чиқиш имконияти;

дастлабки матнлар билан txt, doc и rtf форматда ишлаш, кодлаштиришни автоматик тарзда белгилаш имконияти [Поляков А.Е. <http://www.ruscorpora.ru/new/corpora-biblio.html>].

Шундай қилиб аниқ бўладиган, дастур – конкордансер ёзиш ҳар қандай корпус яратишнинг асосини ташкил этади. Бу корпуснинг жуда катта ҳажмдаги материалларни таҳлил этишга мўлжалланган электрон шакллар ва кенг қўламдаги статистик материалларни йиғишга йўналтирилган ўзига хос жиҳатлари билан боғлиқ.

АДАБИЁТЛАР:

1. Аброскин А. А. Поиск по корпусу: проблемы и методы их решения // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. – 277–282 с.;
2. Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (соғда гаплар мисолида): Филол.фан.бўйича фалсафа доктори (PhD)...дис. афтореф. – Тошкент, 2018.–49 б.;
3. Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (расмий ва илмий услубдаги матнлар таҳрири дастури учун): Филол.фан.бўйича фалсафа доктори (PhD)...дис. афтореф. –Фарғона, 2019.–49 б.;
4. Захаров В.П. Корпусная лингвистика: Учебно-метод. пособие. – СПб., 2005. – 48 с.
5. Касьянов В. Н., Касьянова Е.В. Введение в программирование. – <http://pco.iis.nsk.su/ICP>
6. Касьянова Е.В. Язык программирования Zonnon для платформы .NET // Программные средства и математические основы информатики. – Новосибирск: ИСИ СО РАН, 2004. – С.189–205.
7. Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. –М., 2005.– С.155–174.
8. Mengliev B., Shahabitdinova Sh., Khamroeva Sh., Gulyamova Sh., Botirova A. The morphological analysis and synthesis of word forms in the linguistic analyzer // Linguistica Antverpiensia, 2021, 2021(1), стр. 703-712.
9. Хамроева Ш. Муаллифлик корпусининг муштарак ва ўзига хос жиҳатлари // Сўз санъати халқаро журнали. – №2 (2020) – Б. 80-87. DOI <http://dx.doi.org/10.26739/2181-9297-2020-2>
10. Поляков А.Е. Технология подготовки информации в национальном корпусе русского языка.
11. <http://www.ruscorpora.ru/new/corpora-biblio.html>;
12. Toirova G. The importance of linguistic module forms in the national corpus // Current problems of modern science, education and upbringing (Current problems of modern science, education and upbringing in the region) (Electronic scientific journal), – Urgench. 2020, № 5, –В.155-166. http://khorezmscience.uz/public/archive/2020_5.pdf
13. Тоирова Г. Миллий корпус яратишнинг технологик жараёни хусусида. //Ўзбекистонда хорижий тиллар. Электрон илмий-методик журнал. – Тошкент. 2020, № 2 (31) – Б.57-64. <https://journal.fledu.uz/uz/2-31-2020>
14. Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Филол. фан. бўйича фалсафа доктори (PhD) дис. афтореф. – Қарши, 2018. – 52 б.;
15. Эшмўминов А. Ўзбек тили миллий корпусининг синоним сўзлар базаси: Филол. фан. бўйича фалсафа доктори (PhD) дис. афтореф. – Қарши, 2019.– 45 б.