СЎЗ, МАТН, КОРПУС ЛИНГВИСТИКАСИ: НАЗАРИЯ ВА АМАЛИЁТ

Азиза Шарипова
Муҳаммад ал-Хоразмий номидаги Тошкент ахборот технологиялари университети, фил.ф.д. (PhD), доцент

АННОТАЦИЯ

Мазкур мақолада тилшуносликнинг истиқболли, фаол ривожланаётган соҳаси корпус лингвистикасининг назария ва методикаси ҳақида фикр юритилади. Мақолада маълум бир тарзда ташкил этилган ва элементлари матн бўлган тўплам матн корпусига хам алоҳида урғу берилган. Шунингдек, корпус лингвистикасининг функциялари, усуллари ҳамда босқичларининг хусусиятлари таҳлили ўрганилган. Корпус лингвистикасининг асосий тушунчаси электрон шаклда, тизимли, умумлаштирилган, тил белгилари билан таъминланган ва тил муаммоларини ҳал қилиш учун мўлжалланган, филологик жиҳатдан кенг қамровли лингвистик маълумотлар мажмуаси корпуснинг тавсифлари берилган.

Калит сўзлар: *тил корпуси, лингвистик маркировка, матнлар корпуси, автоматлаштирилган дастурий манба.*

WORD, TEXT, CORPUS LINGUISTICS: THEORY AND PRACTICE

Aziza Sharipova
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Doctor of Philological Sciences (PhD), Associate Professor

ABSTRACT

This article discusses the theory and methodology of corpus linguistics, a promising, actively developing field of linguistics. In the article, special emphasis is also placed on the text corpus, a set of elements that are organized in a certain way and whose elements are text. Also, the analysis of features, methods and stages of corpus linguistics is studied. Moreover, the basic concept of corpus linguistics is the description of a corpus (a philologically comprehensive set of linguistic data in an electronic form, organized, summarized, provided with language symbols and designed to solve language problems) are studied in this article.

Keywords: *language corpus, linguistic markup, corpus of texts, automated software resource.*

СЛОВО, ТЕКСТ, КОРПУСНАЯ ЛИНГВИСТИКА: ТЕОРИЯ И ПРАКТИКА

Азиза Шарипова
Ташкентский университет информационных технологий имени Мухаммада ал-Хоразми, д.ф.н.(PhD), доцент

АННОТАЦИЯ

В данной статье рассматриваются теория и методология корпусной лингвистики, перспективной, активно развивающейся области языкознания. В статье также особое внимание уделяется текстовому корпусу, набору определенным образом организованных элементов, элементами которого является текст. Также проводится анализ особенностей, методов и этапов корпусной лингвистики. Базовым понятием корпусной лингвистики является описание корпуса, филологически всеобъемлющего набора лингвистических данных в электронной форме, организованных, обобщенных, снабженных языковыми символами и предназначенных для решения языковых задач.

Ключевые слова: *языковой корпус, лингвистическая разметка, корпус текстов, автоматизированный программный ресурс.*

КИРИШ

Матнлар корпуси, корпус тилшунослиги деб аталмиш асосий тушунчалардан биридир. Корпус тилшунослиги 1960-йилларда Америка Қўшма Штатларида тилшуносликнинг янги йўналиши пайдо бўлган ва Браун корпуси деб ном олган матнлар корпуси сифатида танилган. Ушбу корпус Америка босма нашрининг лингвистик хусусиятларини акс эттириш учун мўлжалланган эди. У магнит ташувчи (дискета ёки қаттиқ диск)га ёзиб олинган ва умумий ҳажми миллионга яқин сўздан иборат АҚШ босма насрига оид турли матнларнинг беш юз икки минг (502 000) сўз ва сўз бирикмаларини ўз ичига олган.

Браун корпуси бу соҳадаги тадқиқотлар учун катта имкониятлар эшигини очиб берди:

а) бошқа шунга ўхшаш корпусларни яратиш учун ўзига хос стандартга айланди;

б) корпус тилшунослигида янги фаннинг яратилишига туртки бўлди;

с) матнлар корпуси ва корпус тилшунослиги усулларини қўллаш соҳаси корпус яратувчилари кутганидан ҳам анча кенгроқ ва ранг-баранг бўлиб чиқди” [1].

МЕТОДОЛОГИЯ

Бугунги кунда кўпгина дунё тилларининг корпуслари яратилган ва яратилмоқда. Масалан, славян тилларидан, аллақачон чех, поляк, болгар каби

тилларнинг корпуслари мавжуд. Бу борада рус тили корпуси тадқиқотлари ҳам орқада қолмай келмоқда. 1980-йилларда ривожлана бошлаган, 1990-йилларга келиб бироз ривожланишдан тўхтаб қолган бўлса-да, бу йўналиш ҳозирга келиб яна фаол ривожланишни бошлади ва сезиларли натижаларга эришилмоқда, бу ҳақдаги маълумотларнинг аксарияти Интернет тармоқларида мавжуд. [2] Корпус тилшунослиги, В.Захаров таърифига кўра, компьютер технологияларидан фойдаланган ҳолда лингвистик корпусларни (матн корпусларини) қуриш ва улардан фойдаланишнинг умумий тамойилларини ишлаб чиқиш билан шуғулланадиган компьютер тилшунослигининг бир бўлимидир. [3]. М.Копотев ва А.Мустаёкилар ҳам, “Аслида корпуснинг (терминнинг) ўзи икки маънога эга” деб таъкидлашади. Биринчидан, корпусни яратиш назарияси ва методикаси; иккинчидан, корпус тадқиқоти яъни корпус усуллари ёрдамида тил устида изланишлар олиб борилади [2]. Бу барча изланишлар ўзбек тили корпусини ҳам яратишда катта аҳамиятга эга.

Биринчи навбатда корпусларни яратиш назарияси ва методикасини ўрганиш лозим. Корпус тилшунослиги одатда учта асосий усулдан фойдаланади:

1. корпусдан тил ҳақидаги маълумотларни автоматик равишда олиш;
2. ахборотни қайта ишлаш;
3. қайта ишланган маълумотларни текшириш ва талқин қилиш.

Дастлабки икки қадам тўлиқ алгоритмлаштирилган, учинчиси эса ҳозирда мунозарали бўлиб қолмоқда [3].

В.Рыков, ўз навбатида, ишнинг қуйидаги босқичларини ажратиб кўрсатади:

1. нутқий фаолиятнинг тузилишини тақдим этиш зарур;
2. корпусни тузиш учун қандай моддий чекловлар мавжудлигини аниқлаш;
3. матнларни танлаш ва матн корпусини тузиш;
4. корпусни тузиш (компиляция қилиш) [1].

Корпус тилшунослигининг асоси шундаки, тил бутунлай ижтимоий ҳодиса бўлиб, уни тажрибага асосланган маълумотлар, яъни нутқ жараёнида тасвирлаш мумкин. Бу биз сўзловчи ёки тингловчи ўзи айтган ёки эшитган сўзларни, жумлаларни ёки матнларни қанчалик тушунаётганини билмаймиз ёки баъзида билмасликка ҳаракат қилишимизни англатади. Тил қайд этилиши, тавсифланиши ва таҳлил қилиниши мумкин бўлган ижтимоий ҳодиса сифатида матнларда намоён бўлади. [6] Ички, овозсиз матнлар ҳам матндир, лекин уларни кузатиш мумкин эмас ва шунинг учун улар ижтимоий ҳодиса эмас. Матнларнинг аксарияти нутқий фаолият шаклида яъни, жамият аъзоларининг тил ёрдамидаги ўзаро муносабатлари сифатида содир бўлади [3]. Юқорида

айтиб ўтилганидек, корпус лингвистикасининг асосий тушунчаси корпусдир. Корпуснинг бир нечта таърифлари мавжуд. Корпус лингвистикаси инглиз тилида сўзлашувчи мамлакатларда пайдо бўлганлиги сабабли, биз биринчи навбатда инглиз тилида сўзлашадиган илмий муҳитда мавжуд бўлган таърифларни берамиз: *In principle, any collection of more than one text can be called a corpus, (corpus being Latin for "body", hence a corpus is any body of text). But the term "corpus" when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition. The following list describes the four main characteristics of the modern corpus.*

1. Sampling and representativeness
2. Finite size
3. Machine-readable form
4. A standard reference. [4]

Матн корпуси – бу маълум бир тарзда ташкил этилган ва элементлари матн бўлган тўпламдир. Корпусни ташкил этиш уни яратувчиси ёки фойдаланувчисининг прагматик мақсадларига қараб ҳар хил бўлиши мумкин. Корпуснинг таркибий элементлари бўлган матнлар бутун адабий асарни ёки унинг бирон бир қисмини ифодалаши мумкин. Қоида тариқасида, корпус бутунлигича магнитли (машина) ташувчида ёзиб олинган ва бир жойда зич жойлаштирилган деб тахмин қилинади [1].

В.Захаров матнларнинг лингвистик ёки тил корпусини маълум бир лингвистик муаммоларни ҳал қилиш учун мўлжалланган, катта, электрон шаклда тақдим этилган, бирлаштирилган, тузилмали, белгиланган, филологик жиҳатдан кенг қамровли тил маълумотлари тўплами деб тушунади. [3]. Рус тилининг миллий корпусини яратувчилар корпусни қуйидагича таърифлайдилар: Корпус – бу маълум бир тилдаги матнларнинг электрон шаклдаги тўпламига асосланган ахборот-маълумотнома тизими. Миллий корпус айнан ўша тилни мавжудлигини маълум бир босқичида (ёки босқичларида) ва ҳар хил жанрларда, услубларда, ҳудудий ва ижтимоий турлар ва ҳоказоларда ифодалайди.

Бизнинг фикримизча, В.П.Захаровнинг таърифи, ушбу тушунчанинг барча хусусиятларини акс эттирувчи ва уни бир қатор шунга ўхшаш ҳодисалардан (масалан, электрон кутубхоналар) ажратиб кўрсатадиган энг мақбул тарифдир. Шундай қилиб, корпус – бу электрон шаклда тақдим этилган, бирлаштирилган, тузилмали, тил белгилари билан таъминланган ва муайян тил муаммоларини ҳал қилиш учун мўлжалланган катта, филологик жиҳатдан кенг қамровли

лингвистик маълумотлар мажмуасидир. Лингвистик тадқиқот корпуси яхлит восита сифатида икки асосий ташкил қилувчига эга, хусусан:

1. Бевосита маълумотлар массиви (матнлар);
2. Корпус менеджери (ихтисослаштирилган қидирув тизими), бу маркировка (белгилаш ёки аннотация) асосида тадқиқотчи учун зарур бўлган бирликларни бутун маълумотлар мажмуасидан танлаш имконини беради.

Лингвистик маркировка сўзларга махсус кодлар беришни назарда тутади. Кодлар теглар сифатида ҳам танилган (инглизчадан таг - белги), тегларни сўзларга нисбатлаш жараёни мос равишда теглаш (инглизча - таггинг) деб аталади. [3]. Маркировка қанчалик кўп ва хилма-хил бўлса, корпуснинг илмий ва маърифий аҳамияти шунчалик юқори бўлади. Ҳозирги вақтда корпус таркибида бўлиши мумкин бўлган қуйидаги белгилаш (маркировка) турлари шартли равишда ажралиб туради: лингвистик ва экстралингвистик (матнни форматлаш хусусиятлари); муаллиф ҳақидаги маълумотлар (исми, ёши, жинси, ҳаёт йиллари ва бошқалар); ва матн (номи, қайси тилда ёзилган, йили, нашр этилган жойи ва бошқалар). Белгилаш (маркировка)нинг лингвистик турлари орасида қуйидагилар ажралиб туради:

1. Морфологик (гап бўлагининг белгисини, шу бўлакка хос грамматик категория белгиларини ўз ичига олади).
2. Синтактик (синтактик таҳлил натижасидир).
3. Семантик.
4. Анафорик.
5. Просодик (транскрипцияланган товушли нутқ корпусида урғу ва интонацияни тавсифловчи белгилар қўлланилади).

Белгилаш (маркировка) автоматлаштирилган дастурий манбалар ёрдамида амалга оширилади. Белгилаш (маркировка)нинг баъзи бир турлари учун автоматик тизимларни яратиш жуда қийин ва тизим яратиш ишининг асосий қисми қўлда амалга оширилади. Бироқ, морфологик ва синтактик таҳлил учун одатда теглар(таггерс) ва парсерлар(парсерс) деб аталадиган турли хил дастурий воситалар мавжуд.

Бироқ, ушбу тизимларнинг аксарияти қўлда ишлашни талаб қилади, чунки морфологик омонимия ва синтактик ноаниқлик ҳолатларида дастур тадқиқотчига бир нечта ечимларни таклиф қилади, ва тадқиқотчи улардан тўғрисини танлаб олади. Бироқ, янги авлод корпуси ўн миллионлаб сўзларни ўз ичига олади, шунинг учун инсон аралашувини рад этадиган, улар бажариши мумкин бўлган иш ҳажмини мустақил равишда бажарадиган тизимни ривожлантириш тамойили илгари сурилади. Ва ҳатто, чегарасиз бўлиши

мумкин бўлган матнни белгилаш жараёнини тўлиқ автоматлаштириш таклиф этилмоқда. [3] Шундай қилиб, белгилаш жуда кўп вақт талаб қиладиган жараён бўлиб, деярли барча босқичларда, айниқса, белгилаш бирлиги машина "тушунадиган" хусусиятларга эга бўлмаган ҳолларда инсон аралашувини талаб қилади.

А.Шарипова ўз мақоласида тил ўрганишда корпус ва корпус лингвистикасининг умумий эмпирик кўринишини яратишга ҳаракат қилган. Биринчи навбатда тилларни реал контекстда ўрганиш бўлган корпус лингвистикасига таъриф берган. XX-аср бошидан ҳозирги кунгача корпус лингвистикасининг умумий кўриниши тасвирланган. Корпус лингвистикасидан фойдаланиш икки даврда жуда ўхшаш эди, фақат фарқи шундаки, XX-аср бошларида ҳеч қандай компьютер ва технология ишлатилмаган. Шунингдек, корпус турларини, масалан, бир тилли ва параллел корпуслар муҳокамаси қилинган ва корпус адабиётининг ушбу соҳасига киритилган корпус тиллари мисоллари кўриб чиқилган. Корпус лингвистикаси тилшуносликнинг амалий ва тадқиқот соҳаларида умумий тилшуносликнинг асосий йўналишларидан сунъий интеллект ва компьютер лингвистикасига айланганлиги ҳақида батафсил ёритган. [5, 7]

ХУЛОСА

Сўнгги ўн йилликларда компьютер технологияларининг фаол ривожланиши тадқиқот ва таълим жараёнларини оптималлаштириш учун зарур шарт-шароитлар яратди. Компьютер технологиялари, биринчи навбатда веб-технологиялар лингвистика соҳасидаги тадқиқот ва амалий характердаги муаммоларни ҳал қилиш воситаларини ишлаб чиқиш имконини беради. Турли тадқиқот ишлари учун талаб қилинадиган ҳар хил турдаги маълумотларга тез кириш имконияти филолог-олимлар учун янги имкониятлар очади. Махсус тарзда белгиланган маълумотларни танлаш орқали тадқиқот ўтказиш имконини берувчи манбалар – тил корпусига айланди. Улар пайдо бўлиши билан нутқий матнлар билан ишлаш имконияти кенгайди, махсус йўналишда корпус лингвистикаси – тилшуносликнинг матн корпусини ишлаб чиқиш ва улардан фойдаланиш назарияси ва амалиёти билан шуғулланадиган бўлими ажралиб чиқди. Корпуслардан фойдаланишнинг учта асосий йўналиши мавжуд: филологик тадқиқотлар, амалий иш турлари, ўқув жараёни. Тил корпусининг сифати унинг икки компоненти: маълумотлар ҳажми ва турли хил белгилашлар билан аниқланади.

REFERENCES

1. Рыков В. В.Тверской лингвистический меридиан.// Теоретический сборник. – Тверь, 1999. – С. 89-96.
2. Копотев М.В., Мустайоки А. Современная корпусная русистика / М. В. Копотев, А. Мустайоки // Инструментарий русистики: корпусные подходы. Slavica Helsingiensia, 34. Helsinki University Press, 2008. – С. 7-24.
3. Захаров, В.П. Корпусная лингвистика: учебно-методическое пособие / В.
4. П. Захаров. – СПб., 2005.
5. Tony McEnery, Andrew Wilson. Edinburgh University Press, 1996. 206 p.
6. Abdumanapovna, S. A. (2018, October). The contemporary language studies with corpus linguistics. In *Proceedings of the 2nd International Conference on Digital Technology in Education* (pp. 82-85).
7. Ибрагимова, Н. А. (2022). ТИЛШУНОСЛИКДА МАТН ТУШУНЧАСИНИНГ ТАДҚИҚИ ВА ТАҲЛИЛИ. *Oriental renaissance: Innovative, educational, natural and social sciences*, *2*(12), 1299-1304.
8. Bakhronova, D., & Khalikulovna, O. E. (2022). LINGUO-STYLISTIC ANALYSIS OF MEDIA HEADLINES IN ENGLISH AND UZBEK LANGUAGES. *Conferencea*, 5-6.