ЖАҲОН ТИЛШУНОСЛИГИДА КОРПУС ЛИНГВИСТИКАСИ.

Хидиров Отабек Жўрабоевич, ўқитувчи

Жиззах давлат педагогика институти

Аннотация: *Ушбу мақолада жаҳон тилшунослигида корпус ва корпус лингвистикаси ҳамда разметкаланган ва разметкаланмаган корпуслар таҳлил марказига олинган.*

Калит сўзлар:*Корпус, миллий корпус, корпус лингвистикаси, парсинг дастури, синтактик теглаш, теглаш усуллари.*

Корпус ва *корпус лингвистикаси* XX асрнинг 60-йилларида пайдо бўлди. *Корпус* атамаси кўп маъноли бўлиб, тилшуносликда турли объектларни атаб келади. Умуман олганда, корпус − маълум тилга оид материаллар йиғиндиси: у тўлиқ матн ёки унинг катта фрагментини қамраб олади. К.Бауш *корпус* атамасига қуйидагича таъриф беради: “Корпус фақат (ёзма ёки оғзаки) матнлар ёки уларнинг қисмларидан иборат чекланган миқдордаги тил материалидир”[1]. Tuscan Word Centre ходими Джон Синклейрнинг таъкидлашича, корпус тил ёки лингвистик хилма-хилликни тадқиқ этиш, маълумот манбаи сифатида намойиш этиш учун кўринадиган мезонларга мувофиқ танланган электрон кўринишдаги матнлар фрагментидан иборат.

Инглиз адабиётларида разметкаланган ва разметкаланмаган корпуслар фарқланади. Разметкаланмаган корпус – қайта ишланмаган, оддий, “хом” матн, лингвистик информацияга эга бўлмаган корпус. Масалан, present сўзи маълум контекстда от бўлиб келишини кўрсатувчи аниқ факт йўқ. Бу маълумот тушунилади, аммо юзага чиқарилмайди. Разметкаланган корпус ўзида лингвистик ахборот сақлайди: Масалан, present – noun/verb, яъни present сўзи ҳам от, ҳам феъл бўлиб кела олади. Ланкастер университети профессори Ричард Сяо бундай корпуснинг оддий матндан фарқланишини атрофлича шарҳлаб берган[2]. Ўз навбатида, Ланкастер университети

олимлари Тони Макенри, Эндрю Уилсон корпус тилшунослиги (корпус лингвистикаси) тўғрисида қуйидаги фикрларни айтишган: “Корпус лингвистикаси - бу тилни ўрганиш бўлиб, ёзма ёки оғзаки ўқиладиган матнни қайта ишлаш, фойдаланиш ва таҳлил қилиш билан боғлиқ барча жараёнларни ўз ичига олади. Корпус тилшунослиги нисбатан замонавий атама бўлиб, “жонли тил”дан фойдаланишга асосланган методологияга асосланилади. Ҳозирги вақтда корпус тилшунослигининг самарадорлиги ва аҳамиятлилиги компьютер лингвистикасининг ривожланиши билан чамбарчас боғлиқ. Ҳар қандай матн таркибини яратиш тамойиллари унинг амалий мақсадига бевосита боғлиқ[3].

Мутахассислар изоҳланиш (тегланиш) даражасига кўра корпуснинг турли кўринишларини фарқлашади. Қуйида шундай корпус турларига қисқача тўхталамиз.

1. Ихтисослашган корпус − маълум бир турдаги матнлар тўплами: газета матни, илмий мақолалар.
2. Умумий корпус турли хил матнларни ўз ичига олади, матн мазмуни ва жанрига алоҳида талаб ўйилмайди.
3. Қиёсий корпус. Улар турли тилларнинг икки ёки ундан кўп кичик қисмлари, масалан, рус ҳамда немис тили ёхуд битта тилнинг вариантидан иборат. Масалан, Австрия ва Швейцария немис тили версиялари.
4. Параллел корпус турли тиллардаги ўхшаш матнларни ўз ичига олган ички корпусдан иборат. Биринчи корпусдан аслиятдаги матн, иккинчисидан таржима матн ўрин олади.
5. Таълимий корпус − бу чет тилини ўрганаётган шахс учун она тилида сўзлашувчилар томонидан ёзилган матнлар тўплами.
6. Дидактик корпус чет тилини ўқитиш жараёнида фойдаланиладиган тил маълумотларидан иборат.

А.Н.Баранов корпус технологияларига қуйидагича таъриф беради: “Корпус − компютер тилшунослигининг энг муҳим воситаларидан бири.

Улар тилни таҳлил қилиш, матнни лингвистик корпус шаклида тақдим этишнинг амалий вазифаларини ҳал қилишга имкон беради”[4].

М.К.Махмутованинг фикрича, дастлаб, тилларни тадқиқ қилишда корпусдан фойдаланишнинг мақсади турли тил элементларининг частотасини ҳисоблаб чиқишдан иборат эди. Бундай элементлар сўз, сўзшакл, морфема ва иборалар бўлиши мумкин. Корпусдан тил ва нутқ бирликлари бўйича турли хил маълумот ва статистика олиш учун фойдаланиш мумкин[5]. Бундай технология, хусусан, лексикография, сўзни автоматик қайта ишлаш тизимлари соҳасида турли вазифаларни ҳал қилиш имконини беради. Нутқни аниқлаш, синтез қилиш, автоматлаштирилган ва машина таржимаси, имло ва грамматикани текшириш каби мураккаб лингвистик муаммоларни ҳал қилишда статистик усулдан ҳам фойдаланилади. Масалан, корпус материалида қайси сўзнинг турғун иборалар гуруҳига тегишли эканлигини аниқлаш мумкин. Бунинг учун олинган маълумотда, бирликлар ўзаро мунтазам равишда бирика олишини текшириш керак.

Голландиядаги Нижмеген университетида ишлаб чиқилаётган грамматикалар корпус матни ҳолатларида синовдан ўтказилади. Грамматика асосида корпусни қайта ишлайдиган таҳлил дастури тузилади. Олинган ишлов бериш натижалари грамматика маълумотларини қанчалик аниқ тасвирлашини кўрсатади. Шундан келиб чиққан ҳолда, корпус технологияси тилшуносликнинг янги назариялари, сўзни автоматик қайта ишлаш тизимларини синаб кўришга имкон беришини кўрамиз.

1993 йилда Лансастер-Осло/Берген (ЛОБ) корпуси ва Британия миллий корпуси (BNC) яратувчиси Жеффри Леич томонидан 1993 йилда тузилган аннотациялаш постулатлардан бири тил белгиларининг аниқ ва тушунарли тавсифлаш принципи эътиборга молик. Шунингдек, унинг фикрига кўра, умумфойдаланишга мўлжалланган корпуснинг разметкаси учта принципга мувофиқ бўлиши керак.

1. Разметка фойдаланувчи учун қўлланма ёки кўрсатма шаклида мавжуд бўлган таҳлил схемасига асосланган бўлиши, ҳар бир параметр ундан жой олиши керак.

2. Фойдаланувчи учун очиқ корпус разметкаси “назарий жиҳатдан нейтрал” бўлиши лозим: разметка параметрлари барча учун тушунарли бўлган тушунчалар тизимидан ташкил топган бўлиши талаб этилади. Агар корпус аниқ бир лойиҳа учун мўлжалланган бўлса, уни разметкалашда махсус, айнан муаллифга хос, умумқабул қилинган таснифдан фойдаланиш лозим: бунда ҳам тузувчидан у ёки бу тил назариясига таяниш талаб қилинади.

3. Корпус аннотацияси схемаси ким томонидан, қайси аудиторияга мўлжалланганлиги аниқ, равшан кўрсатилиши лозим, чунки корпусдан фойдаланишда юридик ва техник жиҳатдан турли чегаралар мавжуд[6].

Ж.Леичнинг биринчи постулати мукаммал ишланган, аммо у ҳамма корпусда ҳам юзага чиқавермайди. Албатта, барча корпуслар у ёки бу даражада маълумот(тег)лар тизими билан таъминланган бўлади. Ҳар бир параметрнинг қандай маълумот ташиётганлигини аниқлашга доим ҳам эришиб бўлмайди. Бу борада “матнни автоматик қайта ишлаш” (http://www.aot.ru) гуруҳида фаолият олиб борган олимларнинг фаолиятини алоҳида таъкидлаш лозим. Улар асосан, рус тилидаги матнларга автоматик ишлов бериш жараёнларини ишлаб чиқишган; компьютер технологиялари ютуқлари назарий тилшунослик билан бирлаштириб, амалий натижаларга эришилган.

Демак, ҳар қандай синтактик теглар тизимини ишлаб чиқиш учун компьютер технологиялари ютуқлари билан биргаликда ўзбек тилшунослигида синтаксис бўйича яратилган назариялар асосида корпус учун парсер дастури ишлаб чиқиш мумкин. Синтаксис бўйича яратилган барча назарий материалларни тўплаш, ўрганиш, умулаштириш ҳамда синтактик теглар яратишда фойдаланиш зарур.

АДАБИЁТЛАР РЎЙХАТИ

1. Махмутова М.К. Проблемы аннотирования (тагирования) текстов в корпусной лингвистике // Выпускная квалификационная работа ЮУрГУ. – Челябинск, 2018. – 94 с.

2. Xianyao Hu, Richard Xiao. How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis // https://www.researchgate.net/publication/294138492_How_do_English_translations_differ_from_non-translated_English_writings_A_multi-feature_statistical_model_for_linguistic_variation_analysis/references

3. Махмутова М.К. Проблемы аннотирования (тагирования) текстов в корпусной лингвистике // Выпускная квалификационная работа ЮУрГУ. – Челябинск, 2018. – 94 с.

4. Баранов, А. Н. Введение в прикладную лингвистику [Текст]: учебное пособие / А.Н. Баранов. – М.: Изд-во Эдиториал УРСС, 2001. – 347 с.

5. Чернякова, Т. А. Методика формирования навыков студентов на основе лингвистического корпуса [Текст] / Т. А. Чернякова. – Тамбов, 2012. – 149 с.

6. Leech, G. Corpus annotation schemes / G. Leech Literary and Linguistic Computing, 1993. – 8/4. – P. 275-281.