Чориева Умида
ТерДУ ўқитувчиси

Хусусий корпусларни тузишда алоҳида матн устида ишлаш йўллари (А.Қодирий асарлари мисолида)

Дунёда кечаётган интеграция ва глобаллашув жараёнида ўзбек тилини дунёвий тиллар даражасига олиб чиқиш ўзбек тилшунослиги олдидаги муҳим вазифалардан бири ҳисобланади.

Жаҳон миқиёсида табиий тилнинг сунъий интеллект билан узвий муносабатини текшираётган янги йўналишларнинг шаклланиб бўлганлиги ўзбек тилини замонавий ва самарали тадқиқ усуллари орқали илмий ўрганишни талаб қилмоқда. Яъни тилшунос Б.Йўлдошев таъбири билан айтганда, “қуруқ назарий фикрларга тўла тадқиқотлар билан бир қаторда, халқимиз учун фойда келтирадиган компьютер дастурларини ишлаб чиқишда фаол иштирок этиш лозим”.

Жаҳон тилшунослигида алоҳида йўналиш сифатида шаклланаётган корпус лингвистикаси тилшунослигимиз олдига ўзбек тилининг ҳам миллий электрон корпусини яратиш каби долзарб вазифаларни қўймоқда.

Миллий корпусимиз бошқа корпуслардан фарқли равишда ўзбек тилининг барча имкониятлари, маъно бойлиги, асрлар давомида шаклланган луғат заҳираси, бошқа тиллардан устун жиҳатларини ўзида мужассамлаштирган бўлиши лозим. Миллий корпусни яратишдан аввал маълум бир асарнинг хусусий корпусини яратиш лозим бўлади. Бу соҳада тилшунослигимизда бир қанча назарий ишлар қилинган бўлсада, амалий жиҳатдан кичик бир асарнинг ёки матннинг хусусий корпуси ҳали яратилмаган. Табиийки, янгиликнинг тан олиниши қанчалик қийин кечмасин, унинг амалиётга тадбиқи муваффақиятли бўлса, у шунчалик катта ютуқларга эга бўлади.

Тилшунос А.Раҳимов электрон луғатларнинг ишлаш принципларини умумлаштириб, қуйидагича изоҳлаган: Тилнинг ҳар бир сўзига мутаносиб келувчи код ишлаб чиқилади, кодни қайта ишлаш жараёнида зарурий бўлган маълумотлар, таржималар, синоним, антоним ва шарҳларга эга бўлиш мумкин. Мазкур қоида бўйича алоҳида матннинг корпусини тузишда аввало, маълумотлар базаси шакллантирилади, бунда ўзбек тилида яратилган барча бадиий адабиётлар, газета ва журналлардаги мақола ва очерклар, илмий ва сиёсий адабиётларнинг матни лингвистик корпуснинг манбаи ҳисобланади. Масалан, А.Қодирийнинг “Ўткан кунлар” асарининг хусусий корпусини тузиш. Бу жараён босқичма-босқич амалга оширилади.

Аввало, лингвистик корпуснинг маълумотлар омборига матнларнинг электрон варианти жойлаштирилади. Кейинги дастурий таъминот жараёнида матнни моделлаштириш ва алоритмлаш босқичи амалга оширилади. Бунда матндаги сўзларнинг лексик сатҳи ўрганилиб, гуруҳларга ажратилади ва кетма-кетликда жойлаштирилади. Масалан, синоним сўзлар, омоним сўзлар, шевага оид сўзлар, тарихий сўзлар, иборалар, кўп маъноли сўзлар каби. Матндаги ҳар бир сўзни маълум гуруҳларга бўлиб жойлаштиришда сўзнинг услубий бўёқдорлиги, турли маъно нозикликлари, валентлик каби ҳодисаларга алоҳида эътибор бериш лозим. Бу вазиятда ҳар бир сўзни матндаги маъноси бўйича жойлаштирилади.

Лингвистик таъминот босқичида сўзларнинг лексик маънолари изоҳланади. Масалан, сипориш – топшириқ, мишовур – маслаҳатчи, амсоли – ҳоказо, афифа – покиза қиз каби тарихий сўзлар изоҳланади; бўйро, қолапой афзали каби шевага оид сўзларнинг асл маъноси келтирилади; қамчинидан қон томмоқ, чувалган ипнинг учини топмоқ, терисига сиғмай кетмоқ каби ибораларнинг лексик маъноси изоҳланади.

Дастурни тузишда энг мушкул жараён кодлаштириш, яъни матнни инсон тушунадиган тилдан машина (компьютер) тушунадиган тилга ўтказиш

жараёни ҳисобланади. Сўзларни кодлаштириш қуйидагича амалга оширилади. Матндаги сўзларнинг ҳар бири уч қисмдан иборат бўлади:

- 1) сўзнинг тартиб рақами, 2) сўз, 3) код

“Сўзнинг тартиб рақами” (унинг адреси). Дастурий таъминот босқичида ажратилган гуруҳлар ҳар бир сўзнинг адреси ҳисобланади. Масалан, синоним сўзларни биринчи гуруҳ деб олсак, 01С ҳарфи билан ва унинг маънодоши ҳам ушбу кўринишда белгиланади. Шевага оид сўзларни иккинчи деб олсак, 02Ш ҳарфи билан белгиланади ва ҳоказо.

“Сўз” – мутаносиб алфавит ҳарфлари билан белгиланади. Масалан, “гўзал” синоним сўзининг адреси биринчи гуруҳга мансуб бўлганлиги учун 01С деб белгиланади ва кейинги мутаносиб алфавит ҳарфи ёрдамида бош ҳарфи Г бўлганлиги сабабли 01СГ деб кодлаштирилади. Бунда синонимик қатордаги барча сўзлар 01С коди билан ёзилади. Масалан, “гўзал” сўзининг синоними бўлган “чиройли” сўзи 01СЧ кўринишида жойлаштирилади.

“Код” – рақам ва ҳарфлар кетма-кетлиги бўлиб, унда сўз барча зарурий морфологик, синтактик, лексик хусусиятлари ҳамда ушбу сўзнинг қайси сўзга тегишлилиги ҳақидаги маълумотлар жамланган бўлади. Масалан, “гўзал” сўзи учун 01СГ0003 рақамлари билан кодлаштирилади.

Кейинги босқич лемматизация, яъни сўзлик тайёрлаш жараёни. А.Раҳимовнинг “Компьютер лингвистикаси” асарида лемматизацияга қуйидагича таъриф беради: “лемматизация – сўзнинг дастлабки, бошланғич формасини (луғатдаги шаклини – леммасини) ташкиллаштириш техникаси бўлиб, бу жараён ўша сўзнинг бошқа сўз шаклларидан келиб чиққан ҳолда амалга оширилади”.

Бу жараён икки босқичда амалга оширилади: биринчиси, ҳар бир сўзнинг мумкин бўлган барча шакллари белгиланади, иккинчиси, сўз асос ва қўшимчаларга бўлинади. Лемматизация сўзларнинг грамматик валентлиги, қайси аффикслар билан бирика олиш имкониятини ҳам белгилаб беради. Бу жараёнда изланаётган сўзнинг фақат ўзак қисми ажратиб кўрсатилади, унинг

қайси ўринда қандай маънода ишлатилганлиги, омонимик хусусиятлари матн мазмунидан келиб чиққан ҳолда аниқланади.

Кейинги қидирув ва декодлаш, яъни машина тушунадиган тилдан инсон тушунадиган тилга ўтказиш жараёни дастурчи томонидан амалга оширилади.

Дастурнинг ишлаш механизми қуйидагича амалга оширилади:

2. Изланаётган маълум сўз, масалан, "**осмон**" сўзи киритилади, натижада матнда мазкур сўз қайси ўринда, қандай бирикмалар ёрдамида боғланган бўлса барчаси белгиланган ҳолда намоён бўлади. Масалан:

- *...бошида симоби шоҳи салла, устидан қора мовут сирилган совсар пўстин, ичида ўзининг Шамайда тиктиргани* ***осмон****и ранг мовут камзул...*
- *Гўё бу сўзлар Кумушбибининг ўчкан чароғини қайтадан ёқарлар, умид* ***осмон****ининг йўқолиб, яшириниб кеткан юлдузлари яна қайтадан ўз ўринларига келиб қўнғандек бўлурлар...* ва ҳоказо.

3. Кейинги қаторларда киритилган сўзнинг лексик-семантик хусусиятлари (синонимик қатордаги вариантлари, иборалар, ўхшатишлар ва мақоллардаги ишлатилиш ўринлари) келиб чиқади. Масалан, "осмон" сўзи белгиланганда унинг остида "кўк", "само" каби маънодошлари ва асарда улар қатнашган жумлалар чиқади. Масалан,

- *...мен шундай ерни топиб тексам бошим* ***кўк****ка етар еди...*
- *Ерлар ериб, ҳамма ёқ шилт-пилт лой, қўрғон кунгираларидаги қировлар бўғқа айланиб* ***кўк****ка кўтарилмакда едилар.*
- *Юсуфбек ҳожи бошлиқ ер ва* ***кўк****ка сиғмаған музаффар халқ ўрда тевараги ни қуршаб тушди* ва ҳоказо.

4. Матндаги белгиланган сўзлар устига борилганда уларга тегишли маълумотлар чиқади. Масалан,

- *бошим* ***кўк****ка етар еди* (ибора)

- *ер ва **кўк**ка сиғмаған* (ибора) каби. Агар бу сўзлар шевага оид ёки тарихий сўзлар бўлса, уларга тегишли маълумотлар ҳам шу кўринишда намоён бўлади.

Дастурни яна ҳам мукаммаллаштириш учун ҳар бир сўзнинг морфологик келиб чиқиши, синтактик белгиларини ҳам киритиш мумкин. Дастур тилшунос ва луғатшунослардан ташқари адабиётшунослар, муҳаррирлар, ўқитувчи ва журналистлар, ижтимоий фан соҳаси мутахассислари учун манба бўлиш билан бирга турли автоматлаштирилган тизимларни яратишда муҳим аҳамиятга эга.

Миллий корпус лингвистик тадқиқотлар ва тил таълими учун электрон дастурлаштирилган маълумот базаси бўлиб хизмат қилар экан, энг аввало, алоҳида матнларнинг кичик хусусий корпусларини яратиш ва уларни умумлаштириш миллий корпус тузишда пойдевор бўлиб хизмат қилади.

Фойдаланилган адабиётлар рўйхати:

1. А.Раҳимов. Компьютер лингвистикаси асослари. Т.: Академнашр, 2011
2. А.Пўлатов, С.Муҳаммедова, Компьютер лингвистикаси. - Тошкент, 2007.
3. А.Қодирий. Ўткан кунлар. Т.: Ўқитувчи, 1980
4. А.Қодирий. Меҳробдан чаён. Т.: Ўқитувчи, 1982
5. Б.Йўлдошев. Компьютер лингвистикаси: муаммо, вазифва ҳамда истиқбол, Мақола. www.ziyo.uz. sayti
6. Б.Йўлдошев. Компьютер лингвистикаси. – Самарқанд: СамДУ нашри, 2008
7. Ш Сафаров, Б Иулдошев. Компьютер лингвистикасини биласизми? Моҳият. 2004 йил 14 август.